Searching for Biomedical Information on the World Wide Web

R. P. Channing Rodgers
Lister Hill National Center for
Biomedical Communications
Building 38A, Room 9S-916
U.S. National Library of Medicine
8600 Rockville Pike
Bethesda MD 20894
Phone: (301)496-9305
Fax: (301)496-0673
E-mail: rodgers@nlm.nih.gov

Abstract

The rapid growth of biomedical information available via the Internet and its most popular retrieval system, the World Wide Web, has fostered active research and development directed toward locating resources that are appropriate for answering specific queries. The goal is to create tools that optimize information retrieval (as measured by two quantities, precision and recall) while minimizing the effort required by the user. Existing Web retrieval tools can be divided into the following groups: manually maintained topical lists, automatically generated word-based indices, software agents and multi-index searching aids, network cataloging methods, and miscellaneous hybrid and newer approaches. Improvements in current methods should arise from further research into: methods of describing objects on the Web; improved ways of searching for (and within) collections of documents as opposed to single documents; the ability to search for fielded documents; and, ways to describe resources that span intra- and inter-disciplinary as well as cross-cultural linguistic differences. For this last problem, the U.S. National Library of Medicine's Unified Medical Language System (UMLS) will be of great help. As online information retrieval improves, efforts are underway to improve the online information itself; quality control over content is being addressed as the peer-review systems of traditional printed journals migrate into the realm of electronic publication.

Introduction

Although still in its technical (social) adolescence, the World Wide Web (1, 2) has been a phenomenal success by many measures. Little known and used outside of the physics community after its initial release from CERN in 1991, it sprang onto a wider stage in 1993 with the release by the National Center for Supercomputing Applications (NCSA, of the University of Illinois, Urbana-Champaign) of NCSA Mosaic, widely regarded as the first "killer application" for the Internet, which was itself just beginning to leave the largely academic environment in which it had been developed. Since that time, the Internet and WWW have both pursued growth rates (measured in terms of users, affiliated computers, and accessible content) that follow exponential curves. The Web has grown to become the most frequently employed means for humans to access the Internet, so much so that some users confuse the Internet (the network-of-networks that allows the Web to function) with the Web (which is just one of many applications running on the Internet). It has transformed the fundamental paradigm for computer-based information retrieval from one in which the user connects via telephone to a single information provider's vendor-specific (often text based) interface, into one in which the user connects to the Internet through an Internet service provider (ISP) and then employs a single graphical user-interface to fetch multimedia information from multiple information providers.

The key initial idea of the Web's primary inventor, Tim Berners-Lee, was to wed the notions of hypertext and network-distributed information in the simplest possible manner. An exemplar of the *client-server* network computing model, the Web assumes that certain computers (hosts) will be executing *server* software that contains and delivers stored information, and that the user will employ *client* sofware to fetch and display information from various Web servers. Early Web clients (also called browsers) were few in number, ran on a limited number of computing platforms, and were largely text-based. NCSA's contributions to early Web technology include recognition of the value of extending Web protocols to encompass multimedia, and the creation of versions the NCSA Mosaic client that ran on all of the major computer operating systems (Microsoft Windows™, Apple MacOS™, and various versions of UNIX™).

The central technical theme of the Web is *integration*. It is multi-platform multi-media (encompassing text, graphics, audio, and video), and multi-protocol (pulling together multiple independent methodologies for logging into computers remotely, transferring files, and more, into a single unified program). The two underlying technical specifications of the Web, Hypertext Transport Protocol (HTTP, which specifies the method for communicating between server and client) and Hypertext Markup Language (HTML, which provides a method of describing the logical structure and presentational requirements of a document) have proven flexible and extensible.

Another key aspect of the Web is its reliance on *open standards* and *open source*, as opposed to proprietary standards and source. The essential characteristic of an open standard is that it is freely available. Most open standards are also developed and/or elaborated in forums that are readily accessible to interested parties — in the case of the Web, these include the World Wide Web Consortium (W^3C) and the Internet Engineering Task Force (IETF). The availability of open source (freely available source code) has also contributed to rapid evolution of the Web. The source code for Netscape Navigator™, for example, was recently made freely available. The reliance upon open source code has produced a huge outpouring of creative programming activity at sites all over the world.

The success of the Web has had a buoyant effect on underused pre-existing technologies; the Standard Generalized Markup Language (SGML, a method of describing the logical structure of a document), parent to HTML, received renewed attention. The use of

Integrated Services Digital Network service (ISDN, a widely available but little used form of digital telephony) accelerated with growing interest in Web access. Development of a new object-oriented computer language, to be known later as Java, was rescued from possible termination once its developers realized its potential application to the Web.

The Web presents a number of interesting problems: its hypertext links are unidirectional (as discussed below), and the documents they point to can disappear without notice. Users encounter numerous visual design excesses, which will presumably decline as the Web becomes less novel and information providers focus more attention on content. The lack of security adequate for commercial transactions is being addressed by new protocols, as is the issue of exposure of minors to inappropriate content. There are still complaints about delays in accessing information on the Web, although the speed of the Internet increases steadily, as do the connections of individual users to the network.

Among the greatest practical problems for the biomedical user are these: the rudimentary nature of current methods for locating specific content, and the unsystematic application of quality assurance methods such as peer-review. These two problems are coupled, and might be rephrased as one: the difficulty of finding content of adequate quality. This problem is the focus of the present article.

The Organization of Information in the World Wide Web

Many definitions of the term "document" are possible when discussing the World Wide Web, but here it shall be used to mean the content of the display of a Web client, which may include text and separate multimedia objects that have been merged into what appears to the user to be a single entity. The hierarchical organization of information is a well-established convention, as, for example, in the division of a book into chapters, sections, and paragraphs. Thus it is not surprising to find that information available via the Web is often laid out hierarchically. Figure 1A illustrates a hierarchical document, laid out in the form of a family tree. The topmost document might represent a table of contents, the next tier of documents the three chapters in the publication, and so forth. In graph theory and information science, such a structure is referred to as a *tree*. Note that there is one and only one path between any two documents in the tree. A more realistic approximation to the organization of such a document on the World Wide Web appears in Figure 1B, in which the links between documents, instantiated by addresses known as Uniform Resource Locators (URLs), are unidirectional. Mathematically, the structure in Figure 1B is known as a *directed graph*. In this scheme, some documents can not be reached from certain other documents in the set. An important feature of the Web is the capability to create what is in effect a rich set of cross-references, through URLs that point to other sections within a single document and to other documents not directly connected via the hierarchy. Thus the situation is actually as in Figure 1C, in which the documents are richly connected by a mesh of unidirectional links. Cycles and multiple paths from one document to another now become possible, though are not guaranteed for all documents in the set. Formally, the document set is now termed a *network* rather than a simple tree. The Web is in effect a giant networked document set, its rich mesh of unidirectional links between individual documents being analogous to the physical connections between the computers of the Internet, over which most Web traffic travels (though on the Internet, connections are generally bidirectional).

Categories of Web Navigation Aids

Table 1 presents a list of Web navigation systems, categorized by access method. In discussing these methods, an issue that will repeatedly appear is that of *scaling*. A common problem in network programming is that algorithms that work well on local networks, or for small numbers of documents, do not always perform satisfactorily for larger networks or larger numbers of documents.

The earliest attempts to provide systematic access to biomedical Web documents were manually maintained lists of selected URLs, grouped by topic, each document generally described by title. In the hands of a knowledgeable and disciplined reviewer, a list that is focussed on a narrow topic can be a powerful aid to finding information. The use of multiple reviewers (with each reviewer evaluating each potential entry to the list) can help to work against the inevitable biases of an individual reviewer. As the scope of a collection of topical lists expands, multiple selectors/reviewers are required, as it is rare to find one individual with the requisite breadth of knowledge. With multiple persons dividing up the work, however, it becomes more difficult to ensure consistency in evaluations. Manually constructed lists do not scale well to large numbers of documents, and the labor of assembling and maintaining large and complex topical lists can be daunting.

The earliest attempts at automating document location on the Web employed software that recursively retrieves documents pointed to by documents already retrieved, in conjunction with software that creates a word-based index of document content. These retrieval systems are often referred to as *spiders*, *crawlers*, or *Web robots*. These programs require fast network connections, fast computers, and large amounts of storage. These methods have proliferated and are widely used. Their retrieval activities have proved disruptive at some Web sites, and there are conventions which allow a Web server to disallow such indexing from occurring, though adherence to this convention is entirely voluntary on the part of the robot operators. A recent study of six of the most widely used robot-based indexing services(3) estimated that the Web then consisted of 320 million documents (but noted that its true size is almost certainly much larger). Coverage of this document set by the six indexing services that were studied varied from 3% to 34%. Thus, at least two-thirds of available Web content is not referred to by these indices. The number of "broken links" (hypertext pointers that refer to non-available documents) varied from 1.6% to 5.3%. The size of the Web is so vast that even these automated indexing methods have difficulty in scaling to it. This is partially compensated for by the hypertext linking of the Web itself; a document not appearing in the index may well be arrived at via a hypertext link appearing in an indexed document.

Note that most indexing schemes present search results as if the Web were a flat hierarchy of documents under the index, as indicated graphically in Figure 2A. This discards the useful information that is inherent in the hypertext links contained within the documents. Recent work makes use of word-based indices together with hypertext links to assemble topic-based clusters of documents (4). Weights are assigned to documents based on their values as *hubs* (documents that point to other documents) and as *authorities* (documents that are pointed to by other documents), and the weights for a document are adjusted based on the weights of documents pointing to it, and the weights of documents to which it points. This is schematically portrayed in Figure 2B, which shows two topical clusters, and where the weight of a given document is represented by its size.

As vast as the Web is, it may still be but the tip of an information iceberg. Many online biomedical resources exist in the form of pre-Web-era databases. The information in these databases is often richly structured (an important example is the National Library of Medicine's MEDLINE, a bibliographic database in which a major portion of the biomedical literature is indexed according to *fields* such as author, title, publication, date, etc.). Such databases may have forms-based Web interfaces (see Table 2 for a selected list of such interfaces to MEDLINE), but their contents are often invisible to Web indexing engines. An important server-client information retrieval protocol, Z39.50 (5), deals with fielded information sources of this type, but its deployment remains circumscribed, in part due to daunting technical complexity for implementors.

The *hidden tier* of publicly available but non-indexable information may be quite large, perhaps even larger than the current content of the Web, and encompasses many types of information sources, as indicated in Figure 2C. Certain information sources simply do not lend themselves to treatment as text objects. An image database may have written catalog descriptions for its contents, but text is at best an incomplete substitute for an image. Similarly, the textual content of an expert system might suggest the subject area it dealt with, but the values and richness of its information arises in the process of interacting with it; evaluating its text alone is to ignore the decision-making algorithms that constitute its true value. It remains an open question as to whether existing Web-based indexing systems would scale gracefully if suddenly confronted with not just 34% of current Web content, but the complete current content of the World Wide Web, along with all of the other Internet-based information resources lurking within the hidden tier.

Some Simple Explorations With A Popular Retrieval Engine

A systematic formal evaluation of the performance of current Web-based retrieval services is beyond the scope of this presentation, but much can be learned by examining the results from submitting a single biomedical query to one of the more widely used word-based indexing systems, in this case AltaVista. This query attempts to find documents relevant to an actual current medical dilemma: "does estrogen replacement therapy cause breast cancer?" Figure 3 shows the result of naively typing this entire phrase into an AltaVista search form in July 1998 (the lengthy query is truncated in the final display). The system reports finding 1,608,950 matches. When this same search had been done 15 months earlier (in April 1997), AltaVista reported "about 100,000" matches, indicative of the rapid growth of Web-based information (as well as possible evolutions in AltaVista’s content and search software). The system handles the full-text query gracefully, returning a ranked list of URLs. Note also a pictorial advertisement that is relevant to the search topic, as well as a textual advertisement from a book vendor (advertising revenue is a major source of support for commercial indexing services).

Now we examine the ten documents listed in the first display of results from AltaVista. Four of these derive in one way or another from mainstream peer-reviewed medical journals, three derive from an organization that sells products dealing with "life extension," two derive from the medical feature of a U.S. national television news program, and one is a clinical summary written by faculty at a major U.S. medical school.

A closer examination of the URL list reveals that five of the ten documents contain information relevant to the query; however, one of these is simply a "news alert" pointing to another of the five, two are relevant but tied in one way or another to a product being sold by the authoring body, and one is an opinion of a single practitioner, without pointers to original sources. The one remaining URL of this group is a summary of peer-reviewed research, but concerned with patient attitudes rather than the actual risk of breast cancer. One of the ten was partially relevant, but pointed to a collection of literature summaries, one of which (a summary of a review article in a clinical journal) was relevant, but found only after some searching. This collection of summaries was associated with an organization that sells products, and the choice of which summaries to include in the collection may reflect the interests of that organization. Four of the ten were not directly relevant to the original query. A user seeking reports of peer-reviewed research directly relevant to the intent of the query would be disappointed, although continuing beyond the first 10 URLs might reveal more helpful information. Note that the peer-reviewed research information that we did eventually find is present in the form of summaries, not the full text of the original publications.

Using the AltaVista "advanced" search form, entering the boolean construction "estrogen replacement therapy" AND "breast cancer" produced 2,124 matches (the number reported in April 1997 was "about 400"). The first 10 of the resulting URLs present a picture similar to the above: only 2 items were found to be directly relevant to the query, and these pointed to summaries from academic medical centers, without citations of original peer-reviewed research.

This superficial examination of a single search engine does not constitute a rigorous study, but it does highlight certain of the problems confronting the current biomedical searcher. Even carefully phrased searches can yield a huge number of returns, of which only a small fraction are relevant: in terms used by information scientists, the *precision* of the retrieval is low. Much material of value is not retrieved at all, because it is not Web-accessible, or not indexed by the service being used: in the language of information science, the *recall* of the retrieval is low. When recall is high and precision is low, material of value will not be found even if present due to the large size of the retrieval set:

It can take considerable effort to find solid evidence-based information upon which to base serious medical decisions. Much of the peer-reviewed medical literature remains locked away in the hidden tier alluded to earlier, much of it electronically accessible, but, for technical or contractual reasons, not through general Web-based indices. Weeding out commercial pitches and idiosyncratic opinion from solid peer-reviewed research is not an easy or rapid task. Even where apparently solid information is present, its attribution may be unclear and its content poorly documented. The ranking algorithms that various retrieval services use are poorly documented, may change without notice, and are subject to deliberate distortions in the index (there are companies which advertise services intended to help interested parties to modify their document content in ways that will, it is claimed, cause such documents to be placed higher in retrieval rankings for certain types of queries).

Challenges in Building the Global Digital Library

There are several other areas of active research that may contribute to enhanced methods for Web-based information retrieval:

1) Semantic interoperability. The grail of digital library research was succinctly summarized by the editors of a special issue of *IEEE Computer* dedicated to the topic of digital libraries, in the phrase *deep semantic interoperability* (6). Achievement of this goal will require new protocols that can describe, search for, and transport a wide variety of information objects from disparate sites scattered over networks. Semantics is that branch of knowledge concerned with *meaning*, and, the current state of information science is such that meaning is conveyed primarily by language. Semantic interoperability will require that retrieval mechanisms span the inevitable site-to-site variations in the language used to describe comparable content. For biomedicine, there is already in place a powerful aid for semantic bridging: the NLM's Unified Medical Language System (UMLS) (7). The 1998 edition of the UMLS Metathesaurus™ contains 476,322 biomedical concepts, associated with 1,051,903 concept names, drawn from 40 source vocabularies. Several systems for Web-based retrieval have been developed using the Metathesaurus, including an NLM-developed research prototype known as Sourcerer (8) (developed as part of a component of the UMLS project known as the Information Sources Map, or ISM), and the AltaVista-like Medical World Search (see Table 1).

2) Methods for describing objects on the Web. The term *metadata* (9) (data used to describe other data) often arises in this context. Work in various permanent and *ad hoc* bodies such as the World Wide Web Consortium (W^3C) and the Dublin Core Initiative (10) is putting in place standards defining both the means of wrapping and transporting metadata within Web documents, and the types of information that

metadata should contain. There is an emphasis in this work on providing tools that do not require the skills of trained catalogers to use. Tools such as the UMLS Metathesaurus could potentially contribute here, serving as the knowledge base to allow human-assisted or automated cataloging.

3) Methods for choosing from among multiple information resources. Information is often created and managed in discrete collections. Images tend to be gathered and cataloged by experts in images, bibliographic data is created by workers with expertise with bibliographic data, and surgical information is gathered and organized by people with expertise in surgery. An information element from a collection often inherits certain properties from its containing collection (for example, whether or not it has been subjected to peer-review), so it can be useful to know which collection an item comes from. The compartmentalization of information along the lines of subject and data type is not likely to disappear: there are sound institutional, managerial, and quality-control-related reasons for it to continue. However, the simple text-based Web indices of today treat individual information elements across collections, evaluating each element identically (Figure 2A). It would be better to search by recognizing that information is already organized into collections, and devising search methods that use a two- (or more) tiered method, in which appropriate collections are first identified, and then those collections are searched for specific information. This approach would also scale more gracefully as the amount of information to be indexed, and the number of collections, grew. Research continues into ways in which *collections* of information objects can themselves be indexed, so that appropriate collections can be identified in response to a user query (11).

4) Improved ways of dealing with structured (fielded) information. The Z39.50 protocol, mentioned earlier, has much to teach the Web.

Daunting challenges also remain with respect to optimizing resource descriptions, defining the best algorithms for searching for sources, merging and organizing retrieved information obtained from multiple sources (which is often in disparate forms and formats), and in designing the graphical user interfaces that will control the cataloging and retrieval processes.

Equally important is the challenge of successfully migrating the social systems that currently ensure content quality in the print literature, peer-review prominent among them, to the electronic realm. Efforts at quality assurance and peer-review are slowly establishing themselves within the biomedical sector of the Web. Medical World Search (MWS, see Table 1) and Health on the Net Foundation (HON, see Table 1) index only a biomedical subset of total Web content. The process is guided by a peer-review panel of medical professionals in the case of MWS, and HON has promulgated a a list of 8 quality-assurance criteria required for a site to receive "HON Code of Conduct" status.

Ongoing developments suggest that the Web search engine of tomorrow is going to be much more capable than the ones in use today, and that the assessment of the quality of returned information will become easier. These remain early days for the electronic biomedical library.

Acknowledgements

John Kunze (University of California, San Francisco) provided helpful discussions and made numerous suggestions for improvement.

References

1. SCHATZ BR, HARDIN JB. NCSA Mosaic and the World Wide Web: Global hypermedia protocols for the Internet. *Science* 1994; **265**: 895-901.
2. CHANDAK PK, MAMMONE GL, GREENES RA, HOLMAN BL. Internet and the World Wide Web: A Primer for Practitioners. *JMPM* 1995; **11**: 136-142.
3. LAWRENCE S, GILES CL. Searching the World Wide Web. *Science* 1998; **280**: 98-100.
4. CHAKRABARTI S, DOM B, RAGHAVAN P, RAJAGOPALAN S, GIBSON D, KLEINBERG J. Automatic resource compilation by analyzing hyperlink structure and associated text. *Computer Networks and ISDN Systems* 1998; **30**. Proceedings of the Seventh International World-Wide Web Conference (Refer to URL http://www7.conf.au/programme/fullpapers/1898/com1898.html).
5. KUNZE JA, RODGERS RPC. *Z39.50 in a Nutshell (An Introduction to Z39.50)*. National Library of Medicine, July 1995. http://www.lhncbc.nlm.nih.gov/staff/rodgers/publications/z39.50.html.
6. SCHATZ B, CHEN H. Guest Editors' Introduction: Building Large-Scale Digital Libraries. *Computer* 1996; **29**: 22-26. (Refer to the URL http://www.computer.org/computer/).
7. LINDBERG DAB, HUMPHREYS BL, MCCRAY AT. The Unified Medical Language System. *Methods of Information in Medicine*. vol. 32. 1993: 281-291.
8. RODGERS RPC. Automated Retrieval from Multiple Disparate Information Sources: The World Wide Web and the NLM's Sourcerer Project. *JASIS* 1995; **46**: 755-764.
9. CATHRO W. *Metadata: An Overview*. August 1997. Standards Australia Seminar "Matching Discovery and Recovery" (Refer to URL

http://www.nla.gov.au/nla/staffpaper/cathro3.html).

10. Dublin Core MetaData Initiative. http://purl.org/dc/.

11. CALLAN JP, LU Z, INGWERSEN P, FIDEL R. Searching Distributed Collections with Inference Networks. In: FOX EA, ed., *SIGIR '95: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 9-13 July 1995*. Seattle, Washington: 1995: 21-28.

12. SHKLAR L, SHETH A, KASHYAP V, SHAH K. InfoHarness: Use of Automatically Generated Metadata for Search and Retrieval of Heterogeneous Information. *Proceedings of CAiSE-95*. June 1995.

13. DODGE C, MARX B, PFEIFFENBERGER H. Web cataloguing through cache exploitation and steps toward consistency maintenance. *Computer Networks and ISDN Systems* 1995; **27**: 1003-1008. (Paper presented at the Third International World-Wide Web Conference, Darmstadt, 10-14 April 1995).

Table 1. Categories of Web Navigation Aids

Manually maintained lists

Martindale

`http://www-sci.lib.uci.edu/˜martindale/HSGuide.html`

The Argus Clearinghouse

(formerly the Clearinghouse for Subject-Oriented Internet Resource Guides)

`http://www.clkaringhouse.net/`

The World-Wide Web Virtual Library

(subject list maintained by World Wide Web Consortium)

`http://www.w3.org/hypertext/DataSources/bySubject/Overview.html`

Medicine list maintained by Appleyard, OHSU

`http://www.ohsu.edu/cliniweb/wwwvl/`

EINet Galaxy

`http://galaxy.einet.net/`

MedWeb (Emory University)

`http://www.cc.emory.edu/WHSCL/medweb.html`

Planet Earth (Naval Command, Control & Ocean Surveillance Center)

`http://www.nosc.mil/planet_earth/info.html`

Hardin Meta Directory (University of Iowa; a list of lists)

`http://www.lib.uiowa.edu/hardin/md/`

The Medical Matrix (Healthtel Corp.)

`http://www.medmatrix.org/Index.asp`

Table 1 (continued). Categories of Web Navigation Aids
Automated indexing (employing Web crawlers, spiders, software automatons, robots)

World Wide Web Worm (WWWW) (by title and URL; Univ. of Colorado)
    `http://wwww.cs.colorado.edu/wwww`

WebCrawler (full text; Univ. of Washington, America OnLine; allows submissions)
    `http://webcrawler.com/`

Repository-Based Software Engineering (RBSE) JumpStation (full text)
    `http://js.stir.ac.uk/jsbin/js`

Harvest Project (University of Colorado; project now dormant)
    `http://harvest.cs.colorado.edu/`

Lycos (Carnegie Mellon, Lycos, Inc.)
    `http://www.lycos.com/`

AltaVista (DEC; full text)
    `http://altavista.digital.com/`

Excite
    `http://excite.com/`

InfoSeek
    `http://infoseek.com/`

HotBot (WIRED Digital, Inc.)
    `http://hotbot.com/`

Northern Light
    `http://northernlight.com/`

Table 1 (continued). Categories of Web Navigation Aids
Software search agents & multi-index searching aids

SoftBot (Etzioni)
    `http://www.cs.washington.edu/research/softbots`
    underlying technology for the MetaCrawler search service:
    `http://www.metacrawler.com/`

Colorado State's multilingual Savvy Search
    `http://guaraldi.cs.colostate.edu:2000/faq/`

Informarket Search (IBM)
    `http://www.infomkt.ibm.com/`

Search.Com (c| net)
    `http://www.search.com/`

Table 1 (continued). Categories of Web Navigation Aids
Network cataloging efforts

The Mother of All BBS (Univ. of Colorado

`http://wwwmbb.cs.colorado.edu/˜mcbryan/bb/summary.html`

ALIWEB (NEXOR)

`http://web.nexor.co.uk/public/aliweb/aliweb.html`

Catalog-card like descriptions in Summary Object Interchange Format (SOIF)

Gatherer gathers index information locally or across network

Broker provides interface to other brokers or to indices in potentially cascading fashion

Can both replicate and cache data, improving efficiency

Modular, supports arbitrary user lookup interface (default engine is *Glimpse*)

BellCore InfoHarness (object-oriented retrieval based on metadata descriptions) (12)

OCLC, IETF efforts at "metadata" (cataloging) definitions

NLM Sourcerer prototype: cataloging with UMLS knowledge sources (8)

Table 1 (continued). Categories of Web Navigation Aids
Hybrids/Other

Yahoo (submissions; Web crawlers; human organization)

`http://www.yahoo.com/`

HON (Health on the Net Foundation: submissions; Web crawlers; human review)

`http://www.hon.ch/`

Medical World Search (UMLS™-aided word-based indexing of biomedical subset of Web)

`http://www.mwsearch.com/`

HITS/ARC (IBM/Cornell(4); automated indexing with link-weighted clustering)

`http://www7.conf.au/programme/fullpapers/1898/com1898.html`

Google (Document weighting based on how often it is linked to by others)

`http://www.google.com/`

Cataloging of proxy server contents (Dodge et al., 1995(13))

Table 2. Web Access to MEDLINE (Partial List)

PubMed

    http://www.nlm.nih.gov/
    http://www.ncbi.nlm.nih.gov/PubMed/

Internet Grateful Med

    http://www.nlm.nih.gov/
    http://igm.nlm.nih.gov/

Evaluated MEDLINE (from BioMedNet™)

    http://www.biomednet.com/